One-to-one person re-identification for queue time estimation

Lejbølle, Aske Rasch; Krogh, Benjamin; Nasrollahi, Kamal; Moeslund, Thomas B.

Published in:
2020 IEEE International Conference on Image Processing (ICIP)

DOI (link to publication from Publisher):
10.1109/ICIP40778.2020.9191293

Publication date:
2020

Document Version
Accepted author manuscript, peer reviewed version

Citation for published version (APA):
Lejbølle, A. R., Krogh, B., Nasrollahi, K., & Moeslund, T. B. (2020). One-to-one person re-identification for queue time estimation. In *2020 IEEE International Conference on Image Processing (ICIP)* [9191293] IEEE Signal Processing Society. IEEE International Conference on Image Processing (ICIP) https://doi.org/10.1109/ICIP40778.2020.9191293

ONE-TO-ONE PERSON RE-IDENTIFICATION FOR QUEUE TIME ESTIMATION

Aske R. Lejbølle[1,2,*] *Benjamin Krogh*[2] *Kamal Nasrollahi*[1,3] *Thomas Moeslund*[1]

[1]Visual Analysis of People Laboratory, Aalborg University, Denmark
[2]Veovo Denmark, Denmark
[3]Milestone Systems A/S, Denmark

ABSTRACT

Measuring queue times at an airport check-point is essential in order to regulate staff allocation and maintain a low queue time. In this paper, we propose a method to measure queue times based on person re-identification. Specifically, we capture passenger features from entrance and exit points of an airport check-point and match features to determine queue times of passengers. However, a passenger seen at the entrance potentially can be matched to multiple passengers seen at the exit. Therefore, to increase the precision of re-identification, we propose a simple, yet effective, assignment algorithm that assigns each passenger seen at the exit to only a single passenger seen at the entrance. Through experiments on a dataset collected from an airport immigration check-point, we show that our proposed assignment increases precision and recall by 22 % and 16 %, respectively, to naively assigning the best match.

***Index Terms*—** Re-identification, Assignment problem, Hungarian algorithm

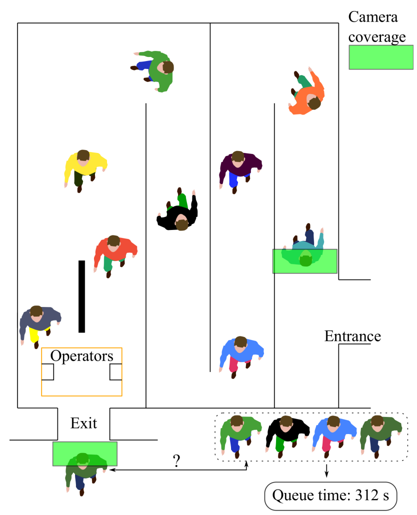


Fig. 1: Principle of re-id in a queue. Passengers follow a predefined maze and are captured by cameras both at the entrance and exit. Features captured by the exit camera are matched against all passengers features captured by the entrance camera to find a match and output a corresponding queue time.

1. INTRODUCTION

The task of matching characteristics, i.e., features, from persons captured across non-overlapping cameras in a camera network is also known as person re-identification (re-id). Person re-id is often linked to forensics where an operator inputs an image of a person (probe) to a system, which matches the image against a database of known persons (gallery), and returns a list of the most likely matches. In this case, the re-id problem is defined as an image retrieval problem [1]. Furthermore, this case often considers a closed-world setting where the same persons have appeared in all cameras. On the other hand, there are scenarios, such as trajectory tracking of multiple persons across a large camera network [2], where re-id is considered as a verification task and the probe captured by one camera is individually matched against all gallery persons. In this setting, we have no knowledge about the appearance of a probe in the other cameras, therefore, we cannot simply consider a list of likely matches. This is also known as an open-world setting [3]. In this work, we consider the closed-world setting, however, without an operator to observe the returned list. Instead, we wish to automatically identify matches to a given set of probes that are correct with a high probability. In this case, we might consider additional contextual information, such as the number of times a person in the gallery can be matched to a probe.

This work aims to find a set of correctly matched persons from a gallery in a closed-world setting using additional contextual information. More specifically, we consider a queue in an airport, where features of passengers captured by an overhead camera at the exit are matched against those captured by an overhead camera at the entrance, to identify matches and return the corresponding time that passengers spent in the queue, as shown in Figure 1. Between the entrance and exit, the queue might temporarily split, thus, while we do know that the same passengers appeared at both entrance and exit, we cannot naively assume a first-in-first-out scenario. In-

*Corresponding author.

stead, we slack the assumption and base our assignment on a likelihood of the order being consistent. Additionally, we make a valid assumption that each person in the gallery can be matched to only a single probe.

We propose to solve this problem as an optimization problem, that is, we wish to reduce the total matching cost between the gallery and a set of probes. In re-id, the likelihood of a match is defined by a distance of some pre-defined metric, therefore, minimizing the total cost corresponds to minimizing the total distance between the probes and gallery. We propose to apply the Hungarian algorithm [4] to solve the assignment problem and increase precision. In re-id, the Hungarian algorithm has previously been used to minimize the cost of matching image patches across probe and gallery [5, 6], but has not yet been applied in a post processing step to increase the re-id precision. After applying the assignment step, we are able to more accurately measure queue times of each passenger. To summarize, we provide the following contributions:

- We propose the use of person re-id to measure queue times in an airport. To our knowledge, this is the first time re-id has been reported to be used for this purpose.
- We propose a post processing step to increase the precision of re-id using the Hungarian algorithm to minimize the total distance between probes and gallery.
- Through experiments, we show that re-id based queue time measurements can produce median queue times that are close to ground truth.

2. RELATED WORK

Re-ranking: To our knowledge, this is the first work to perform optimized probe-gallery assignments as a post processing step to a re-id system. Previous work [7, 8, 9] focus on re-ranking of the initial output to increase precision. Early re-ranking methods consider the k-nearest neighbors of a probe to produce new ranked lists [7, 10]. Other methods consider reciprocal neighbors to generate new features that are matched using a Jaccard distance, and either re-rank based on the new distance [8, 11] or fuse the Jaccard and original distance before re-ranking [12]. More recently, Sarfraz et al. [9] proposed an expanded cross neighborhood (ECN) re-ranking by aggregating distances from nearest neighbors and reciprocal nearest neighbors to avoid the cost of encoding of new features. Different from re-ranking, which still allows a gallery person to be matched to multiple probes, our assignment algorithm adds the constraint that each gallery person can be assigned to only a single probe.

Queue time measurements: Current queue time measurements systems nowadays either utilize data captured from WiFi/Bluetooth (BT) devices to track passengers throughout the queue [13] or track passengers using newly deployed cameras [14]. While WiFi/BT device solutions are cheap, they are challenged by future changes in the operating systems of mobile devices. Camera based solutions that track passengers do not suffer from this issue, however, in low ceiling areas a potentially large number of cameras is required to cover the entire queue. Meanwhile, a re-id solution, as in this paper, only requires two cameras; one at the entrance and another at the exit.

3. METHODOLOGY

An overview of the proposed methodology is shown in Figure 2. We extract features from probes and gallery and perform feature matching. The resulting distance matrix is used as input to the assignment algorithm, which assigns the id's of gallery persons to each probe based on minimizing the total distance between probes and gallery.

In this work, we use an existing convolution neural network (CNN) architecture, which was developed to perform re-id from an overhead viewpoint [15]. The CNN is a multimodal architecture, which processes RGB and depth images in parallel using a MobileNetV2 as backbone [16], and fuses modality features in late layers of the network. For each modality, the network contains soft attention mechanisms [17] to capture and dynamically weight local semantics that are fused with global feature representations. Finally, the two modality based features are fused to a multimodal feature representation containing fused global and local information. Please see [15] for a more detailed description of the architecture.

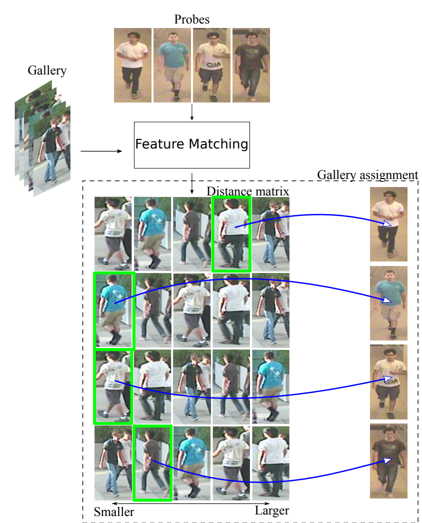


Fig. 2: Overview of the assignment procedure. Each probe is matched against all persons in the gallery to produce initial lists that are ranked by distance, while green boxes indicate true matches. Afterwards, the Hungarian algorithm [4] is applied to assign each gallery to only one of the probes by minimizing the total distance.

We apply the CNN to extract features $f_i^p, f_j^g \in S$ from probes and gallery, respectively, where f_i^p is the feature descriptor from the *i*'th probe, and f_j^g is the feature descriptor from the *j*'th gallery, furthermore, $\{f_i^p, f_j^g\} \in R^{128}$. Using probe and gallery features, we calculate the Euclidean distance $D_E(f_i^p, f_j^g)$ between the i'th probe and j'th gallery features. We further consider the order of which passengers entered the queue by adding an additional distance, based on assigning entrance and exit id's to each passenger. The id's are assigned based on the order of which passengers entered and exited the queue, respectively. Note that, due to an intermediate queue split, the same passenger might have different entrance and exit id's. We define the entrance and exit id's as $enter_j$ and $exit_i$, respectively, and calculate an order distance between the i'th probe and j'th gallery as $D_O(enter_j^g, exit_i^p) = \log(1 + |enter_j^g - exit_i^p|)$. We take the logarithmic value of the distance, since we can have potentially large values compared to Euclidean distances, depending on the number of persons in the dataset. Hence, each probe and gallery id contain a set $x_i^p = \{f_i^p, exit_i^p\}, x_j^g = \{f_j^g, enter_j^g\}$ of both a feature vector and exit or enter id.

Next, we apply the Hungarian algorithm to find an optimal set S^* of matches that solve the following optimization problem:

$$\begin{aligned} \mathcal{S}^* =& \arg\min_{\mathcal{S}} \sum_{x_i^p, x_j^g \in \mathcal{S}} D_E(f_i^p, f_j^g) + D_O(enter_j^g, exit_i^p) \\ &\text{Subject to} \quad x_i^p \neq x_m^p, x_j^g \neq x_n^g \\ &\forall \{x_i^p, x_j^g\}, \{x_m^p, x_n^g\} \in S, \end{aligned} \tag{1}$$

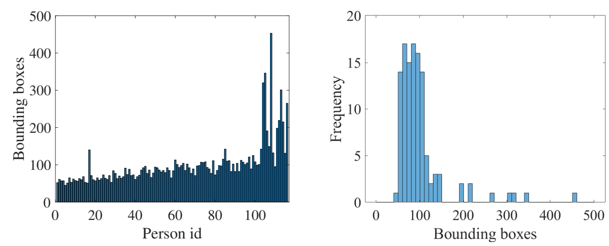
where the optimal set, S^*, is the one where the sum of distances between features and id's in the set S is minimized and $\{x_i^p, x_j^g\}, \{x_m^p, x_n^g\}$ are assigned pairs. The constraint dictates that each of the id's in an assigned pair cannot be assigned to any other id.

4. EXPERIMENTS

To evaluate the performance of re-id for queue time measurements, we evaluate the proposed algorithm on a re-id dataset containing timestamps. To our knowledge, only the public dataset of [18] contains timestamps, however, the dataset was collected in an outdoor environment at a university campus, which does not comply with our goal of using timestamps to measure queue times. As a result, we have collected and annotated a new dataset we call *Queue Person Re-identification* (QPR). In the following, the dataset is briefly introduced along with the experimental results of our proposed assignment algorithm and a comparison to related state-of-the-art methods.

4.1. QPR Dataset

We collect data from an immigration area at an airport to properly evaluate queue time measurements using person re-id. Data is collected using two ZED cameras [19] that are placed overhead at non-overlapping locations. The cameras are placed at the entrance and exit of the immigration area, respectively, to capture the queue times of each passenger. We extract RGB images and compute disparity maps using semi-global block matching (SGBM). Disparity maps are converted to depth maps, and a JET color map is applied to create depth images that can be used along color images to train the CNN, as each pixel in the depth image represents depth as RGB values. We annotate bounding boxes around all persons in the dataset, resulting in 7529 bounding boxes across 116 persons. An overview of the dataset statistics is shown in Figure 3. A few persons, primarily those that entered the queue late, have more than 200 annotated bounding boxes, nonetheless, the majority of persons in the dataset have [50,100] annotated bounding boxes, as seen in Figure 3 (b).



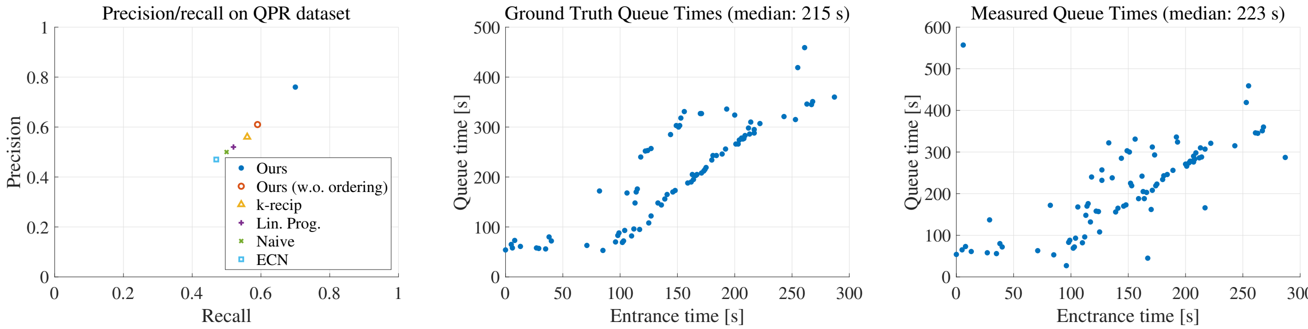
(a) (b)

Fig. 3: (a) Number of bounding boxes per person and (b) frequency in number of bounding boxes in bins of 10.

4.2. Implementation Details

As mentioned in Section 3, we apply an existing CNN architecture [15]. The QPR dataset is split between 26 persons that are used for training and 90 that are used for testing. The network is trained using a combination of cross-entropy and triplet loss with soft margin using hard positives and negatives [20], which has been a popular approach to train CNNs [21, 22]. We use Adam optimization [23] to update the weights of the CNN, with a base learning rate of 0.0001, which is reduced by a factor of 0.9 every 100 epochs, and decay values of 0.9 and 0.999, respectively. Finally, we train the network using a batch size of 32 until convergence, but up to 1000 epochs. For each batch, we randomly sample four images from eight randomly sampled persons.

Upon testing, we extract features from all images of each person, and perform average pooling on the features to create a single feature descriptor for each person in each view. We match feature descriptors between the two views using Euclidean distance, and correspondingly, calculate the order



(a) (b) (c)

Fig. 4: (a) Comparison of precision/recall values to related state-of-the-art re-ranking, (b) ground truth queue times and (c) measured queue times using re-id.

distances between probes and gallery. The output distance matrices are then used as input to our assignment algorithm. To measure the performance of our system, we report precision and recall values.

4.3. Experimental Results

To identify the number of most likely matches to consider in our assignment, we perform various experiments where we consider different numbers of most likely matches when assigning gallery to probes, that is, we consider the top-k gallery matches. The results of our proposed algorithm using different values of k are shown in Table 1. If we naively assign the most likely match, we achieve a precision and recall of 50 %. If we apply our assignment algorithm and just consider the most likely match, we see a large increase in precision, however, at the cost of a lower recall. Nonetheless, considering the six most likely matches, precision and recall increase to values that are 22 % and 16 % higher, respectively, compared to the naive approach. Finally, we see that precision and recall converge if 10 or more most likely matches are considered.

	Naive	Top-1	Top-3	Top-6	Top-10	Top-20	Top-50	Top-90
Precision [%]	50	71	65	76	67	66	66	66
Recall [%]	50	46	57	70	67	66	66	66

Table 1: Precision and recall of our proposed assignment algorithm by considering 1, 3, 6, 10, 20, 50 and 90 most likely matches. As reference, we compare to naively assigning by taking the most similar match.

We further compare our results to two additional baselines; (1) applying our assignment algorithm without the order distance and (2) by applying linear programming [24] (Lin. Prog.). We also compare our results with the k-reciprocal (k-recip) re-ranking of [12] and the ECN re-ranking of [9] using their publicly available code. The results are shown in Figure 4 (a). It is clear that our method outperforms all baseline and re-ranking methods that either do not consider the order of passengers or perform optimized one-to-one assignments. Interestingly, ECN performs worse compared to Naive. The reason hereof might be the direct aggregation of reciprocal distances, which is more effective in a multi-shot case, where the probability of reciprocal neighbors similar to that of the probe is higher. In this work, we only consider the single-shot case.

To further highlight the effectiveness of our assignment algorithm, we perform profiling on a laptop containing an Intel i7 @ 2.60 GHz CPU. Running the algorithm for 1000 iteration, we calculate a mean processing time of 2.11 ms.

Finally, we apply our assignment to measure queue times and calculate a median queue time, as shown in Figure 4 (c). For comparison, we also plot the ground truth queue times, as shown in Figure 4 (b). In both cases, entrance time of the first passenger is 0 s. As seen in both plots, similar tendencies are followed with slightly more scatter in case of the re-id based queue times. The differences in queue times result in a median difference of 8 seconds, which is a deviation of only 3.60 %.

5. CONCLUSION

We have proposed person re-identification to measure queue times in an airport. To increase re-id precision, we have proposed a one-to-one assignment algorithm by applying the Hungarian method to minimize the total distance between a gallery and a set of probes, which considers both the Euclidean distance and an order distance, based on the order of which passengers entered and left the queue. We have evaluated our algorithm on a novel overhead RGB-D person re-id dataset, which is collected from a queue scenario in an airport and contains passenger timestamps. Through experimental results, we have shown that the proposed assignment algorithm can increase re-id precision and recall by up to 22 % and 16 %, respectively, which results in a median queue time that deviate 3.60 % from ground truth.

6. REFERENCES

[1] Liang Zheng, Shengjin Wang, Lu Tian, Fei He, Ziqiong Liu, and Qi Tian, "Query-adaptive late fusion for image search and person re-identification," in *Proc. CVPR*, 2015, pp. 1741–1750.

[2] Siyu Tang, Mykhaylo Andriluka, Bjoern Andres, and Bernt Schiele, "Multiple people tracking by lifted multicut and person re-identification," in *Proc. CVPR*, 2017, pp. 3539–3548.

[3] Qingming Leng, Mang Ye, and Qi Tian, "A survey of open-world person re-identification," *IEEE Trans. on Circuits and Systems for Video Technology*, 2019.

[4] Harold W Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[5] Weiyao Lin, Yang Shen, Junchi Yan, Mingliang Xu, Jianxin Wu, Jingdong Wang, and Ke Lu, "Learning correspondence structures for person re-identification," *IEEE Trans. on Image Processing*, vol. 26, no. 5, pp. 2438–2453, 2017.

[6] Slawomir Bak and Peter Carr, "One-shot metric learning for person re-identification," in *Proc. CVPR*, 2017, pp. 2990–2999.

[7] Xiaohui Shen, Zhe Lin, Jonathan Brandt, Shai Avidan, and Ying Wu, "Object retrieval and localization with spatially-constrained similarity measure and k-nn reranking," in *Proc. CVPR*, 2012, pp. 3013–3020.

[8] Herve Jegou, Hedi Harzallah, and Cordelia Schmid, "A contextual dissimilarity measure for accurate and efficient image search," in *Proc. CVPR*, 2007, pp. 1–8.

[9] M. Saquib Sarfraz, Arne Schumann, Andreas Eberle, and Rainer Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking," in *Proc. CVPR*, 2018, pp. 420–429.

[10] Mang Ye, Jun Chen, Qingming Leng, Chao Liang, Zheng Wang, and Kaimin Sun, "Coupled-view based ranking optimization for person re-identification," in *Proc. MMM*, 2015, pp. 105–117.

[11] Danfeng Qin, Stephan Gammeter, Lukas Bossard, Till Quack, and Luc Van Gool, "Hello neighbor: Accurate object retrieval with k-reciprocal nearest neighbors," in *Proc. CVPR*, 2011, pp. 777–784.

[12] Zhun Zhong, Liang Zheng, Donglin Cao, and Shaozi Li, "Re-ranking person re-identification with k-reciprocal encoding," in *Proc. CVPR*, 2017, pp. 1318–1327.

[13] Veovo, "Unleash the power of predictive insights," `https://veovo.com/platform/passenger-predictability/`, 2019, Accessed: November 27, 2019.

[14] Xovis, "Security checkpoint," `https://www.xovis.com/solutions/detail/security-checkpoint/`, 2018, Accessed: November 27, 2019.

[15] Aske R Lejbølle, Kamal Nasrollahi, Benjamin Krogh, and Thomas B Moeslund, "Person re-identification using spatial and layer-wise attention," *IEEE Trans. on Information Forensics and Security*, vol. 15, no. 1, pp. 1216–1231, 2019.

[16] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. CVPR*, 2018, pp. 4510–4520.

[17] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[18] Meng Zheng, Srikrishna Karanam, and Richard J Radke, "Rpifield: A new dataset for temporally evaluating person re-identification," in *Proc. CVPR Workshops*, 2018, pp. 1893–1895.

[19] Stereolabs, "Zed - depth sensing and camera tracking," `https://www.stereolabs.com/zed/`, 2017, Accessed: November 28, 2019.

[20] Alexander Hermans, Lucas Beyer, and Bastian Leibe, "In defense of the triplet loss for person reidentification," *arXiv preprint arXiv:1703.07737*, 2017.

[21] Jianlou Si, Honggang Zhang, Chun-Guang Li, Jason Kuen, Xiangfei Kong, Alex C. Kot, and Gang Wang, "Dual attention matching network for context-aware feature sequence based person re-identification," in *Proc. CVPR*, 2018, pp. 5363–5372.

[22] Jifei Song, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales, "Generalizable person reidentification by domain-invariant mapping network," in *Proc. CVPR*, 2019, pp. 719–728.

[23] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[24] Jirí Matouek and Bernd Gärtner, *Understanding and Using Linear Programming (Universitext)*, Springer-Verlag, Berlin, Heidelberg, 2006.